IDENTIFICATION OF MEDICAL DOCUMENTS

BY

LANCELOT HOGBEN, F.R.S.

MURIEL M. JOHNSTONE

AND

K. W. CROSS

(From the Department of Medical Statistics, University of Birmingham and Birmingham United Hospital)

To fulfil administrative requirements for cross-reference with respect to successive admissions of one and the same person to the same medical institution and to different ones, or to subserve the ends of research involving reference to original records, a system of medical documentation must make provision for identification of individual patients. *Ipsa facto* the method of designation must be highly specific in the sense that it excludes the possibility of confusing two individuals. It must also be suitable for transcription by recourse to a numerical code. Otherwise it is not feasible to take advantage of current mechanical methods of sorting and tabulation involving the use of punch-cards.

For the purposes of the punch-card, numerical coding of the birth name is unsatisfactory both because it monopolizes a considerable fraction of space available for other uses and because allocation of requisite columns is highly variable. The complete birth name of an individual may require any number of letters, from a minimum of about seven—e.g., May Hall—to a maximum of about 30. For instance, Henry William Howard Waterhouse would entail the allocation of 28 columns by the alpha system of punching and sorting with respect to a single letter of the alphabet per column.

Frequency of Surnames

Equally important, though too rarely recognized, is the fact that birth names are not highly specific. Farr recognized this long ago, and emphasized its statistical implications in words which are still worthy of citation: "From the indexes of the registers for the year 1853, the probable number of persons in England and Wales bearing each of the 50 most frequent surnames has been computed. . . . The persons by whom these 50 surnames are borne amount to about 3,253,800; nearly one-sixth of the entire population of England and Wales. On an average, it seems, one person in 73 is a Smith, one in 76 a Jones, one in 115 a Williams, one in 148 a Taylor, one in 162 a Davies, and one in 174 a Brown."

In an investigation into the frequency of surnames in a Midland telephone directory we found that the following seven names accounted for 5.6% of 88,351 persons listed: Smith, 1,330 (1.5%); Jones, 965 (1.1%); Williams, 614 (0.70%); Taylor, 550 (0.62%); Green, 520 (0.59%); Brown, 502 (0.57%); Evans, 468 (0.53%).

In one year (1946) the number of new in- and out-patients treated in the Birmingham United Hospital was also about 88,000. With such an intake we might therefore expect the central file of the hospital to accumulate in the course of 10 years records of several thousand Smiths, Joneses, Williamses, Taylors, Greens, Browns, and Evanses. Evidently the task of extracting the records of an individual from a set of documents filed by birth name, or extracting the file number of an individual from a ledger of birth

names, would assume alarming dimensions if there were centralization of medical documents in an administrative region.

In the long run, filing by serial code number without recourse to a ledger of birth names is therefore an administrative necessity. It is also the only practicable system of operating large-scale therapeutic trials or comparable statistical inquiries which entail recourse to detailed information not recorded on the punch-card case summary. Given a punch-card file based on diagnosis, it is possible by present methods of mechanical sorting and tabulation to run off a nominal roll (see Truelove, S. C., and Hogben, L., *Brit. J. soc. Med.*, 1947, 1, 18) of such serial numbers to identify the location of the individual in a central file of complete medical documents with minimum effort.

Three Requirements of Coding

To subserve both ends stated above, a system of serial code numbers must conform to three requirements. It must ensure that: (a) two patients will rarely receive the same serial number; (b) the *fixed* number of ciphers employed do not make excessive demands on the available number of columns of a standard punch-card; and (c) record officers in different institutions will assign without difficulty the same serial number to one and the same patient.

The objective defined by (a) is to ensure the possibility of turning up a patient's records with the utmost dispatch. Since the complete records will necessarily contain information sufficient to distinguish occasional individuals who receive the same serial number, it is not necessary to insist on the absolute specificity of the serial number assigned to him or her. All that matters is to circumscribe the search for a particular document within narrowly defined limits. As regards (b) the allocation of 9 or 10 columns is certainly not excessive, being less than the mean number requisite to code a full birth name as such and on all fours with the new national registration numbers of three letters followed by seven ciphers. The remaining criterion (c) calls for more detailed comment. It carries with it two desiderata: (1) individual ciphers must be referable to personal information which is easily accessible to the record officer and liable to no dubiety with respect to its interpretation; and (2) such information must refer to such characteristics of the individual as do not change in his or her lifetime.

That the choice of such information must be consistent with the fulfilment of the criteria of specificity and of economy of column space imposes further limitations. In particular, it excludes sex as a single unit of classification because the specification of the two sexes leaves vacant eight out of the ten rows of a column assigned to it. It excludes birth rank for a less obvious reason. Though it would be easy to assign the birth rank to a single column by reserving the last row for tenth or later live births in the patient's sibship, such an arrangement would not be economical inasmuch as high-grade specificity presumes *equipartition of statistical frequency to each of a set of exclusive classes*. As it happens, more than one-third of all children are firstborn. Hence the probability that two individuals taken from a population at random will have the same birth rank is far greater than 0.1, which is the minimum attainable for a single column of 10 rows.

The desideratum specified by (1) above excludes numerical coding of blood groups because this would necessitate the performance of a laboratory test, and of birthplace because this would entail recourse to a voluminous gazetteer. Birthplace is also exceptionable on account of (2), because topographical landmarks often change in the lifetime of an

individual consequent upon redistribution of local government responsibilities. For both reasons we may also dismiss fingerprints, since their interpretation is too specialized an art, and the loss of a limb destroys the possibility of subsequent identification. Aside from this, fingerprints have associations which do not find favour with a considerable body of patients. It goes without saying that eye colour, marital status, religion, and occupation are each inconsistent with (2) above, though eminently appropriate if (1) were the only prerequisite.

Considerations stated above are largely relevant to any satisfactory method of serial-number specification attached to a birth certificate or analogous personal document for purposes other than the one which is the theme of this discussion. Admittedly, a quite arbitrary national identity number assigned at birth would dispense with the need for an *ad hoc* examination of the issue in this context; but we have as yet no guarantee that the wartime system of national identity numbers will survive, and any proposal for its permanent adoption would encounter a formidable body of public opinion suspicious of regimentation. In any case, the present system is for several reasons unsatisfactory as it stands. One is that individuals on leaving the armed Services at present receive an identity number which is not the one they had before enlistment. Another, which will be more evident at a later stage, is that it lacks the merit of making explicit information, which has otherwise to be coded separately for general use.

Six-cipher Code

The principle implicit in the last statement offers an opportunity for illustrating how it is possible to make use of a run of ciphers embodying information which is at the same time (a) immediately accessible to a recording officer, (b) invariant with respect to the lifetime of the individual, and (c) useful for a wide variety of statistical purposes. A hospital record must give the age of the patient, and the six ciphers which can specify this suffice to make the datum explicit on any document which bears a serial number in which they occupy a *fixed* position. For instance, the birth date of one of us is Dec. 9, 1895, that of another Aug. 16, 1925. Transcribed direct by day, month, and year without specification of the century, as in general irrelevant, these would be respectively 091295 and 160825. Clearly the act of assigning such a label to an individual involves no issue of regimentation, since no one except the record officer and hospital staff with access to the confidential documents involved would need to know the patient's number. Nor would the patient need to know it. Equally clearly, record officers in different institutions would fill in the same run of ciphers assigned to the same fixed position in the serial number in the same way. These six digits do not necessarily require six columns on a punch-card. Since there are only 12 calendar months, it would be possible to record the two ciphers standing for the month on one column by recourse to the X and Y pigeon-holes.

We have now to ask what level of specificity a six-cipher sequence based on the birth date would guarantee. A convenient measure is the chance that two individuals taken at random will have the same cipher sequence. The chance that any individual will have the same birthday as another is of the order of 1 in 360. In Britain we may put the chance that two individuals taken at random will be of the same age—i.e., that they will have been born in the same year—at about 1 in 50. Thus the number of live individuals in England and Wales with one and the same six-cipher birth-date sequence should be about 2,000.

Ten-cipher Code

By employing four additional ciphers we can greatly reduce the possibility of confusion: how greatly depends on partition of information employed to assure approximately equal relevance to each of the 10,000 compartments which four ten-row columns accommodate. Whether such partition is more or less easy is an empirical issue. Accordingly we have made an analysis of the initial run of letters in 88,000 surnames of a Midland telephone directory to explore the possibility of grouping such initials in 100 blocks of equal frequency with a two-figure code for each block—e.g.:

AA—AK	... 00	BAA—BAJ	... 04
AL	... 01	BAK—BAQ	... 05
AM—AR	... 02	BAR	... 06

If p_x is the proportion of surnames assignable to the x th block we may define a convenient coefficient of specificity (C_s) for the two relevant columns in accordance with the following definition:

$$C_s = \frac{x = 99}{\sum_{x=0} P_x^2}$$

As against the theoretically realizable C_s (=0.01) a straightforward code designed as stated on the basis of information supplied by an 88,000 sample of surnames yielded the highly gratifying figure 0.01073. This result does not justify us in offering for national use the specimen surname initial code (C) printed below, since the best code for the whole nation should be based on a national sample more representative of the population than the contents of a regional telephone directory. None the less, the Midland region is fairly representative of the country as a whole, and we may therefore assert with confidence that a foolproof two-cipher surname code devised for use on the *same pattern* would prove to have a C_s not far off 0.01. The only name in our sample which requires a block to itself is Smith (1.5%). If Smiths are more common in the country as a whole than they are in the Midland region it will be possible to deal with them as we deal below with first names such as William.

To make good use of two additional ciphers we have examined the possibility of coding first names on a principle which makes the sex of the individual explicit in the first digit. For this purpose we have prepared a frequency distribution of 3,000 *male* first names and 4,000 *female* first names in our hospital register. The result shows that certain first names have a relative frequency much higher than 0.01. Where this is so it is possible to ensure approximate equipartition of information by recourse to live-birth rank specification—e.g.:

	Code No.
Francis	15
Frank	15
Frederick, Birth rank 1	16
" " > 1	17
F (excluding above)	14

By assigning the two-cipher sequence 00–49 (Code A) to males and 50–99 (Code B) to females we are able to distribute first names with a C_s equal to 0.0103. As it stands we do not claim that such a code is ideal for the nation, or even for the Midland region. We are content to assert that Codes A and B provide a pattern for a more satisfactory attempt in conformity with the requirements stated above.

The complete proposal which emerges from the foregoing analysis of surnames and first names undertaken to

test the possibility of devising a system of serial numbers easy to record and having a high level of specificity is as follows:

Frederick Henry Allinson, born May 12, 1917, and his twin sister Mary Alice, with one elder sib, would thus be respectively: 1701120517 and 8701120517. Two manifest advantages of this code spring to the eye. For all records on which the serial number appears, the sex of the individual is explicit in virtue of the first digit (0-4 if male, 5-9 if female) and the current age is apparent from the last two ciphers. What is equally important is that the requisite three nominal codes are each short enough to print on one side of a standard card; and the specification of the individual by recourse to the code entails access to no information of a type other than such as the recording officer has to obtain in the normal routine of completing an admission card.

The choice of the word "type" in the last sentence is deliberate, because the code numbers for the third and fourth digits refer to the *birth* surname—i.e., maiden name in the case of a married woman; but any system of filing by surname must also rely on the surname of the individual *at birth* if it is to ensure filing of documents referable to one and the same individual in one and the same place. The problem of legal adoption admits of a satisfactory solution, but need not concern us in this context, since legal adoption is a comparatively rare contingency and usually takes place at a very early age. Co-twins of unlike sex necessarily receive a distinctive first cipher, and are labelled as such by the circumstance that their serials have the same sequence of the residual nine. Co-twins of like sex with first names which have the same initial letter will usually be distinguishable, because first names with the same initial, if also common, will *de facto* have distinctive code numbers. In any case, it is no disadvantage to have the records of two twins in the same niche of a file. On the contrary, it simplifies the task of finding the records of a co-twin for research which brings into focus the nature-nurture issue.

A satisfactory figure for the specificity of the 10-cipher (nine-column) code as a whole would have to give due weight to what correlation exists between the three blocks of data. Some correlation there must be, since it is highly probable that Alberts were more common among male babies at the time of the Great Exhibition than they are to-day. In this matter history is happily on our side. Film-star notoriety rather than goodness seems to be the prevailing fashion, and film-star first names are less prone to supersaturate the already sufficiently large classes of our A and B codes than the Williams, Alberts, Edwards, Georges, Elizabeths, and Marys on which royalty rings the changes.

It is pertinent to remark that what correlations do exist do not necessarily favour confusion. Against the fact that father and son, necessarily with the same third and fourth code ciphers, are somewhat apt to have the same first name,

Code A, for Males Only

×	×								
Albert	Birth rank	1	00	Job	..	..	..	32	
"	"	>1	01	J (excluding above)	..	..	..	26	
Alfred	..	..	02	K	..	..	..	26	
Arthur	Birth rank	1	03	Leonard	..	..	..	33	
"	"	>1	04	Leslie	..	..	..	33	
A (excluding above)	..	..	05	L (excluding above)	..	..	..	26	
B	..	..	06	M	..	..	..	34	
Charles	Birth rank	1	07	N	..	..	..	34	
"	"	>1	08	O	..	..	..	35	
C (excluding Charles)	..	..	09	Percival	..	..	..	34	
Dais	..	..	10	Philip	..	..	..	36	
David	..	..	10	P (excluding above)	..	..	..	35	
D (excluding above)	..	..	11	Q	..	..	..	35	
Edward	..	..	12	Reginald	..	..	..	34	
Ernest	..	..	13	Richard	..	..	..	37	
E (excluding above)	..	..	14	Robert	..	..	..	34	
Francis	..	..	15	R (excluding above)	..	..	..	39	
Frank	..	..	15	Samuel	..	..	..	40	
Frederick	Birth rank	1	16	Sidney	..	..	..	41	
"	"	>1	17	Stanley	..	..	..	41	
F (excluding above)	..	..	14	S (excluding above)	..	..	..	42	
George	Birth rank	1	18	Thomas	Birth rank	1	42		
"	"	>1	19	"	"	>1	44		
G (excluding George)	..	..	20	T (excluding Thomas)	..	..	42		
Harold	..	..	21	U	..	..	42		
Harry	..	..	22	V	..	..	42		
Henry	..	..	23	Walter	..	..	43		
Herbert	..	..	24	William	Birth rank	1	46		
H (excluding above)	..	..	25	"	"	2	47		
I	..	..	26	"	"	3	48		
James	Birth rank	1	27	"	"	>3	48		
"	"	>1	28	W (excluding above)	..	..	42		
John	"	1	29	X	..	..	42		
"	"	2	30	Y	..	..	42		
"	"	>2	31	Z	..	..	42		
Joseph	..	..	32						

and hence the same first and second code ciphers, we have to weigh the fact that they necessarily have a different specification with respect to the last two ciphers of the serial. Admittedly, the last two ciphers take no stock of centenarians or near centenarians, but any danger of confusion arising from this limitation is trivial. Taking all circumstances into consideration, we have good reason to suppose that intercorrelations between the three basic blocks of data embodied in the code specification will not grossly reduce a measure of specificity computed in accordance with an assumption of independence.

Without recourse to an algebraic analysis of the problem—this will be dealt with in a forthcoming publication—we may regard the issue of specificity in the following way. No two Smiths can come in the same niche of a filing system unless born on the same day of the same year. At the present time Smith is the most common patronymic, and the number of Smiths born in Britain per day is about 30. Of these 30 Smiths 15 will be separated by the initial cipher of the serial number in the female half of the file from the remaining 15 males. The chance that two of the 15 Smiths of the same sex will have the same specification with respect to the first two ciphers of the serial is about 1 in 50. Evidently, therefore, two Smiths will rarely get the same serial number or occupy a close position in a national filing system containing records of half a million Smiths.

One minor issue which calls for comment is the occurrence of first names, such as Evelyn or Francis, which are not distinctively male or female. The proposed system takes this into account by prescribing a separate code A for the

Code C. Initial Letters of Surname at Birth

		X	X						
--	--	---	---	--	--	--	--	--	--

AA-AK	..	00	GO-GQ	..	35	PEO-PH	..	68
AL	..	01	GRA-GRE	..	36	PI-PN	..	69
AM-AR	..	02	GRF-GZ	..	37	PO-PRH	..	70
AS-AZ	..	03	HAA-HAL	..	38	PRI-PZ	..	71
BAA-BAJ	..	04	HAM-HAQ	..	39	QA-QZ	..	97
BAK-BAQ	..	05	HAR	..	40	RA-RG	..	72
BAR	..	06	HAS-HAZ	..	41	RH-RI	..	73
BAS-BD	..	07	HB-HH	..	42	RJ-ROB	..	74
BEA-BEK	..	08	HIA-HIL	..	43	ROC-RZ	..	75
BEL-BEZ	..	09	HIM-HOK	..	44	SA-SB	..	76
BF-BJ	..	10	HOL	..	45	SC-SHA	..	77
BK-BN	..	11	HOM-HOU	..	46	SHB-SHZ	..	78
BO-BQ	..	12	HOV-HUF	..	47	SI-SM	..	79
BRA-BRN	..	13	HUG-HZ	..	48	(excl. Smith)		
BRO-BRZ	..	14	IA-IZ	..	97	SMITH	..	80
BS-BZ	..	15	JA-JD	..	49	SN-SS	..	81
CA-CG	..	16	JE-JOM	..	50	STA-STN	..	82
CH-CK	..	17	JON-JZ	..	51	STO-SZ	..	83
CL-CN	..	18	KA-KH	..	52	TA-TD	..	84
COA-COL	..	19	KI-KZ	..	53	TE-TH	..	85
COM-COP	..	20	LA-LD	..	54	TI-TO	..	86
COQ-COZ	..	21	LE-LH	..	55	TP-TZ	..	87
CP-CZ	..	22	LI-LN	..	56	UA-UZ	..	98
DAA-DAU	..	23	LO-LZ	..	57	VA-VZ	..	99
DAV-DAZ	..	24	MAA-MAN	..	58	WAA-WAQ	..	88
DB-DI	..	25	MAO-MAR	..	59	WAR-WAZ	..	89
DJ-DZ	..	26	MAS-MD	..	60	WB-WG	..	90
EA-EL	..	27	ME-MI	..	61	WH	..	91
EM-EZ	..	28	MJ-MOQ	..	62	WIA-WILK	..	92
FA-FH	..	29	MOR-MZ	..	63	WIL	..	93
FI-FL	..	30	NA-NE	..	64	WILM-WIZ	..	94
FM-FO	..	31	NF-NZ	..	65	WJ-WOO	..	95
FP-FZ	..	32	OA-OZ	..	97	WOP-WZ	..	96
GA-GH	..	33	PAA-PAR	..	66	X-Z	..	98
GI-GN	..	34	PAS-PEN	..	67			

A delegation from Charing Cross Hospital visited five hospitals in Europe to study methods of hospital design and planning. The five hospitals were the Neue Bürgerspital in Basle; the Polyclinic in Zurich; the Kinderspital in Zurich; the Beaujon Hospital in Paris; and the New Southern Hospital in Stockholm. This visit was arranged by the King Edward's Hospital Fund for London, and the long report which was subsequently prepared by Sir John Stewart-Wallace, Mr. Norman C. Lake, Dr. H. W. C. Vines, and Dr. A. Doayne Bell was reviewed in the *Journal* of Nov. 1, 1947 (p. 704). The report has now been published and copies are available, price 1s., from the Honorary Secretaries of the Fund at 10, Old Jewry, London, E.C.2.